|  |  |
| --- | --- |
|  | **PROJECT 2 SPRING 2018:**  **KOBE BRYANT SHOT SELECTION !!!** |

**OVERVIEW:**

Kobe Bryant marked his retirement from basketball by scoring 60 points in his final game as a member of the Los Angeles Laker team on Wednesday, April 12, 2016. Starting to play professional basketball at the age of 17, Kobe earned the sport’s highest accolades throughout his long career. Using 20 years of data on Kobe's shots made and shots missed, can you predict which shots will be successful?

**DATA:**

The original data set contains the location and circumstances of every shot attempted by Bryant during his 20-year career. Your task is to predict whether the basket went in (shot\_made\_flag = 1) or missed (shot\_made\_flag = 0). The data for estimation is in project2Data.xlsx.

For this exercise, 5000 of the shot\_made\_flags have been removed from the original data set and are shown as missing values in the project2Pred.xlsx file. These are the test set shots for which you must submit a classification. You are provided a sample classification file, project2Pred.xlsx with the shot\_ids needed for your predicted classification. Provide you predicted classifications in this file and submit both your paper and the prediction file. I have the actual values of the shot\_made\_flag for these missing shot\_ids and will evaluate the classifications. Your goal is to provide the best predictions possible.

Each group is on the honor system to not use any information outside of the dataset to predict each of the missing shot flags.

**DATA CONTINUED**

The field names are given below (Data descriptions are available in Kaggle):

|  |  |
| --- | --- |
| action\_type  combined\_shot\_type  game\_event\_id  game\_id  lat – court location identifier (latitude)  loc\_x - court location identifier (x/y axis)  loc\_y- court location identifier (x / y axis)  lon - court location identifier (longitude)  minutes\_remaining – (in period)  period  playoffs  season  seconds\_remaining  attendance  avgnoisedb – avg noise in arena (decibels) | shot\_distance  shot\_made\_flag (this is what you are predicting)  shot\_type  shot\_zone\_area  shot\_zone\_basic  shot\_zone\_range  team\_id  team\_name  game\_date  matchup  opponent  shot\_id  arena\_temp (oF) |

**DELIVERABLE:**

Students will submit a paper with an 8 page limit with a separate Appendix up to 5 pages. Code should be in a second appendix and can be as long as necessary. A separate file with predicted classifications also should be submitted.

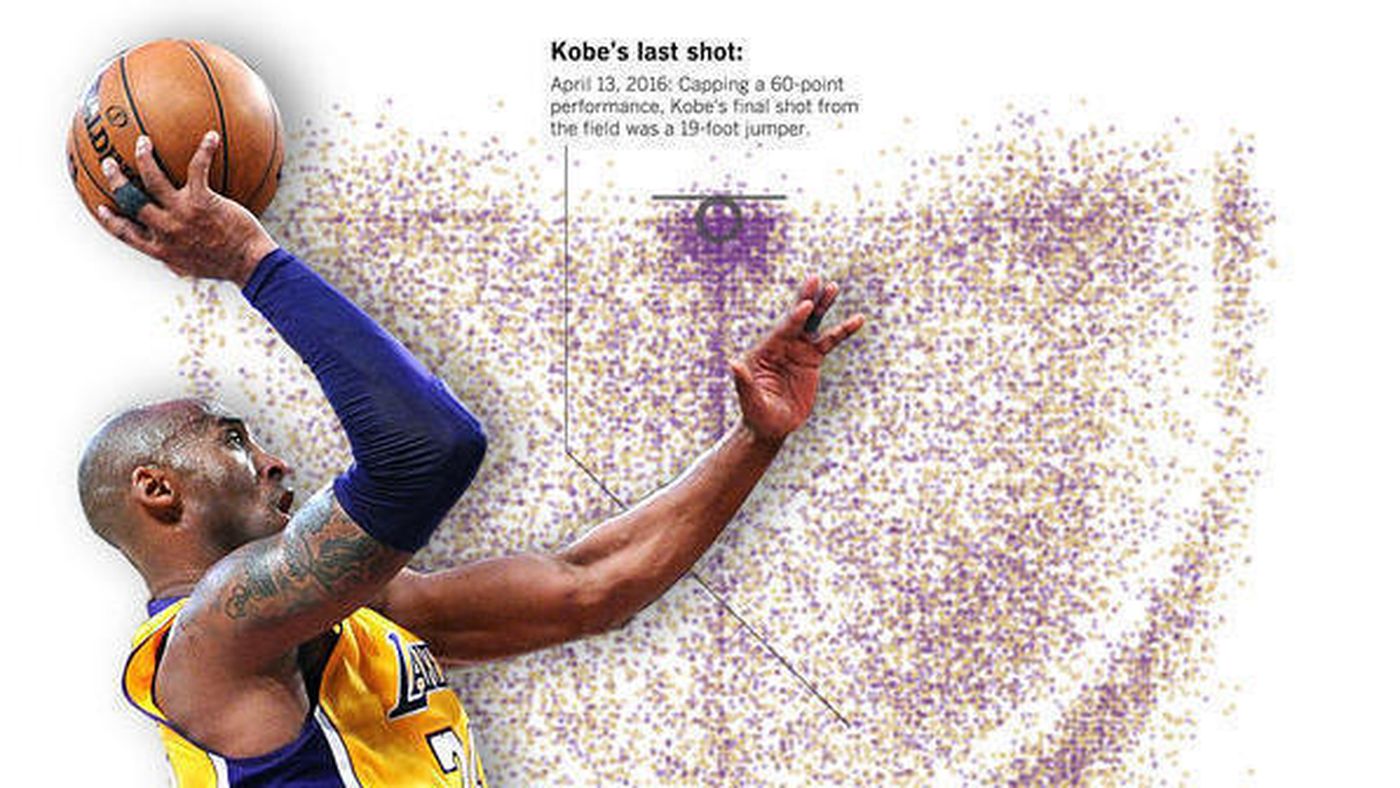
**Kobe Bryant Shot Selection**

**MSDS 6372 - Project 2**

**Fall 2018**

**Allen Crane**

**Brock Friedrich**



**Introduction**

Kobe Bryant marked his retirement from basketball by scoring 60 points in his final game as a member of the Los Angeles Lakers team on Wednesday, April 13, 2016. Starting to play professional basketball at the age of 17, Kobe earned the sport’s highest accolades throughout his long career. Using 20 years of data on Kobe's shots made and shots missed, we wish to predict which shots will be successful.

**Data Description**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Alphabetic List of Variables and Attributes** | | | | |
| **#** | **Variable** | **Type** | **Format** | **Description** |
| 2 | action\_type | Char | 17 | The type of shot that was attempted (e.g. "Jump Shot", "Driving Dunk") |
| 28 | arena\_temp | Num | BEST12. | The average temperature in the arena during the game |
| 27 | attendance | Num | BEST12. | The number of people attending the game |
| 29 | avgnoisedb | Num | BEST12. | The average noise (in decibels) during the game |
| 3 | combined\_shot\_type | Char | 9 | The type of combined shot that was attempted (e.g. "Jump Shot", "Dunk") |
| 23 | game\_date | Num | MMDDYY10. | The date that the game occurred on |
| 4 | game\_event\_id | Num | BEST12. | The id of the specific event associated with this shot attempt |
| 5 | game\_id | Num | BEST12. | The id of the specific game |
| 6 | lat | Num | BEST12. | The location (in latitude) of the shot attempt |
| 7 | loc\_x | Num | BEST12. | The x coordinate of the shot attempt (based on the x y grid of the court) |
| 8 | loc\_y | Num | BEST12. | The y coordinate of the shot attempt (based on the x y grid of the court) |
| 9 | lon | Num | BEST12. | The location (in longitude) of the shot attempt |
| 24 | matchup | Char | 11 | The matchup of the two teams playing (e.g. "LAL @ POR") |
| 10 | minutes\_remaining | Num | BEST12. | The number of minutes remaining in the period in which the shot attempt occurred |
| 25 | opponent | Char | 3 | The NBA abbreviation of the opposing team |
| 11 | period | Num | BEST12. | The period of the game in which the shot attempt occurred |
| 12 | playoffs | Num | BEST12. | Whether or not the game was a playoff game |
| 1 | recId | Num | BEST12. | The unique id of the observation of the shot attempt (in chronological order) |
| 13 | season | Num | YYYYYY | The YearYr indicator of the game season (e.g. "200001" = 2000-2001 season) |
| 14 | seconds\_remaining | Num | BEST12. | The number of seconds remaining in the period (concatenated with associated minutes)  in which the shot attempt occurred |
| 15 | shot\_distance | Num | BEST12. | The distance of the shot attempt (in feet) |
| 26 | shot\_id | Num | BEST12. | The id of the specific shot attempt within the game |
| 16 | shot\_made\_flag | Num | BEST12. | Whether or not the shot attempt resulted in a successful shot |
| 17 | shot\_type | Char | 14 | The technical term for the type of shot that was attempted (e.g. "2PT Field Goal") |
| 18 | shot\_zone\_area | Char | 21 | The zone location on the court that the shot attempt occurred ("Left Side(L)", "Center(C)") |
| 19 | shot\_zone\_basic | Char | 21 | The approximate zone location on the court that the shot attempt occurred ("Mid Range",  "Restricted Area") |
| 20 | shot\_zone\_range | Char | 15 | The distance range in feet of the zone location to the goal ("8-16 ft.", "16-24 ft.") |
| 21 | team\_id | Num | BEST12. | NBA league team id |
| 22 | team\_name | Char | 18 | NBA league team name |

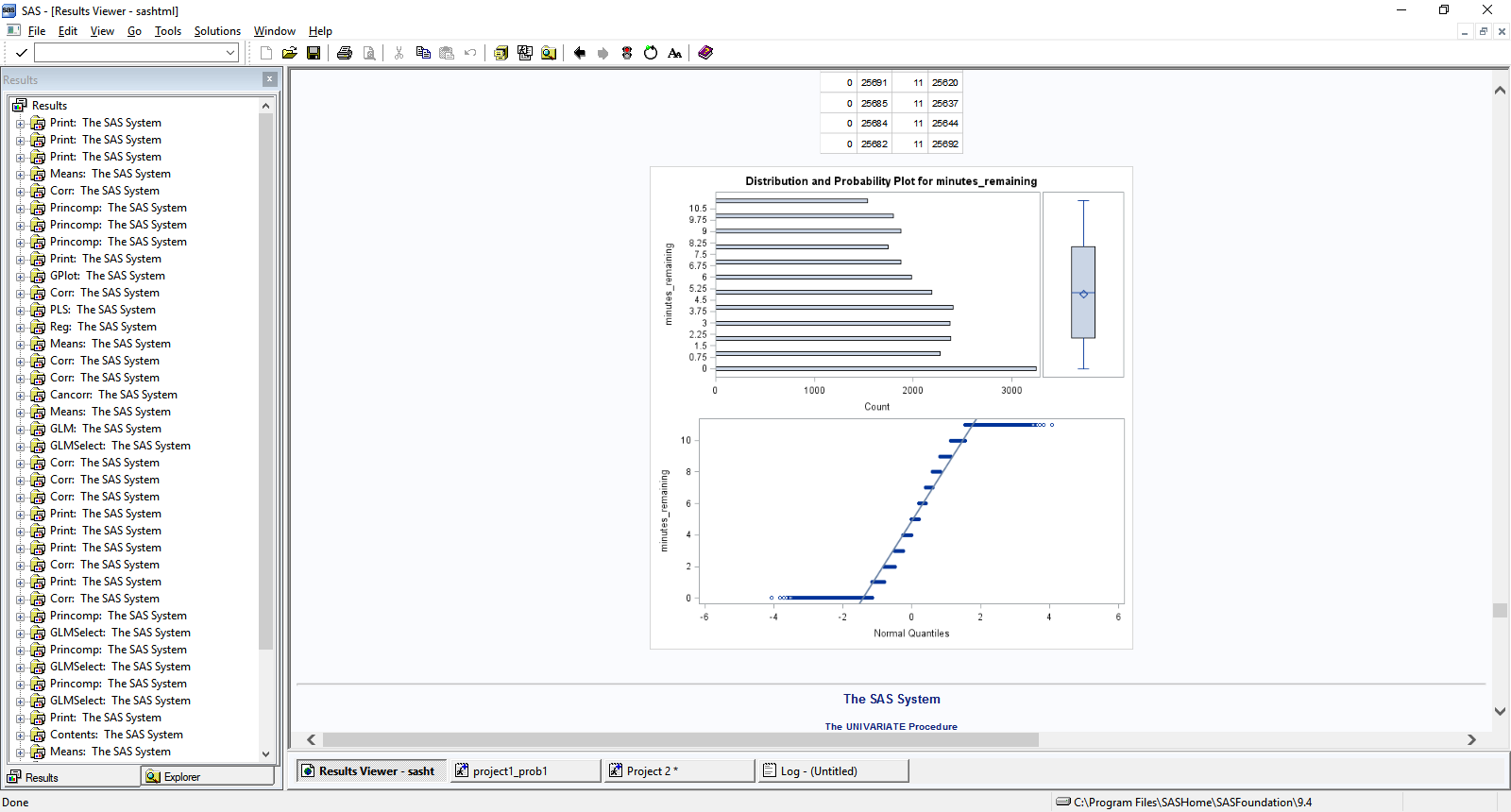
**Exploratory Data Analysis**

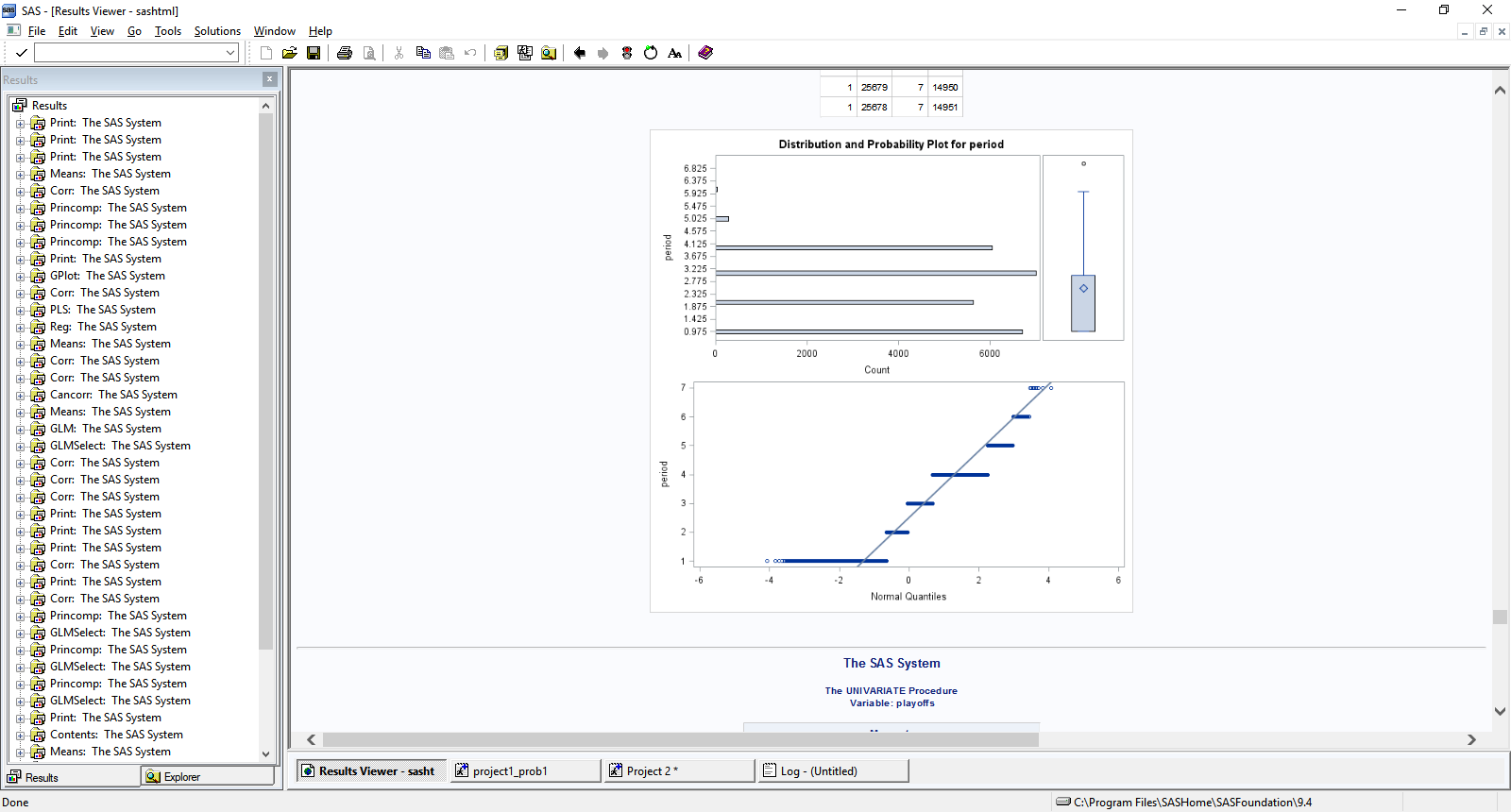
* Address the need for any potential transformations

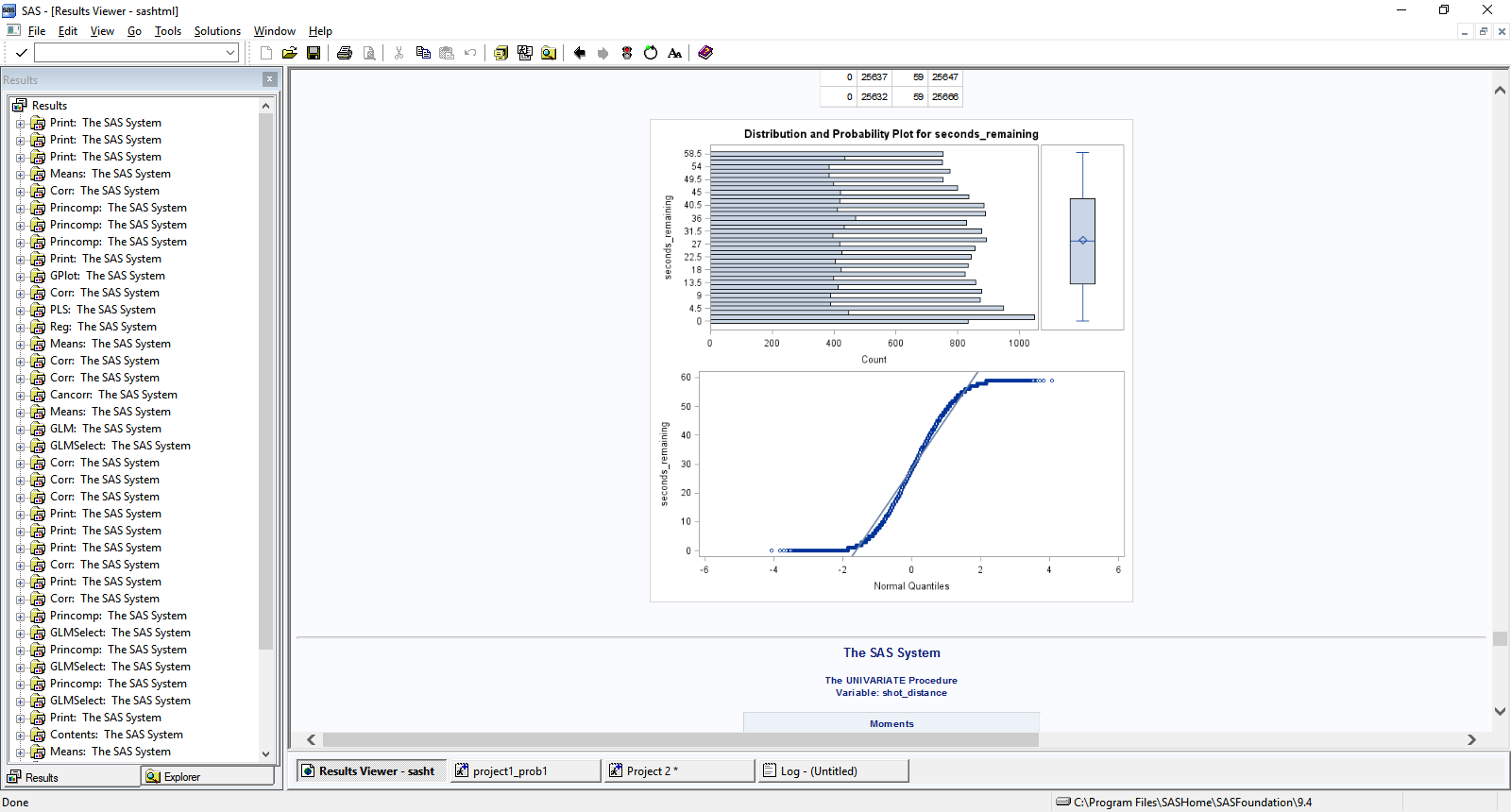
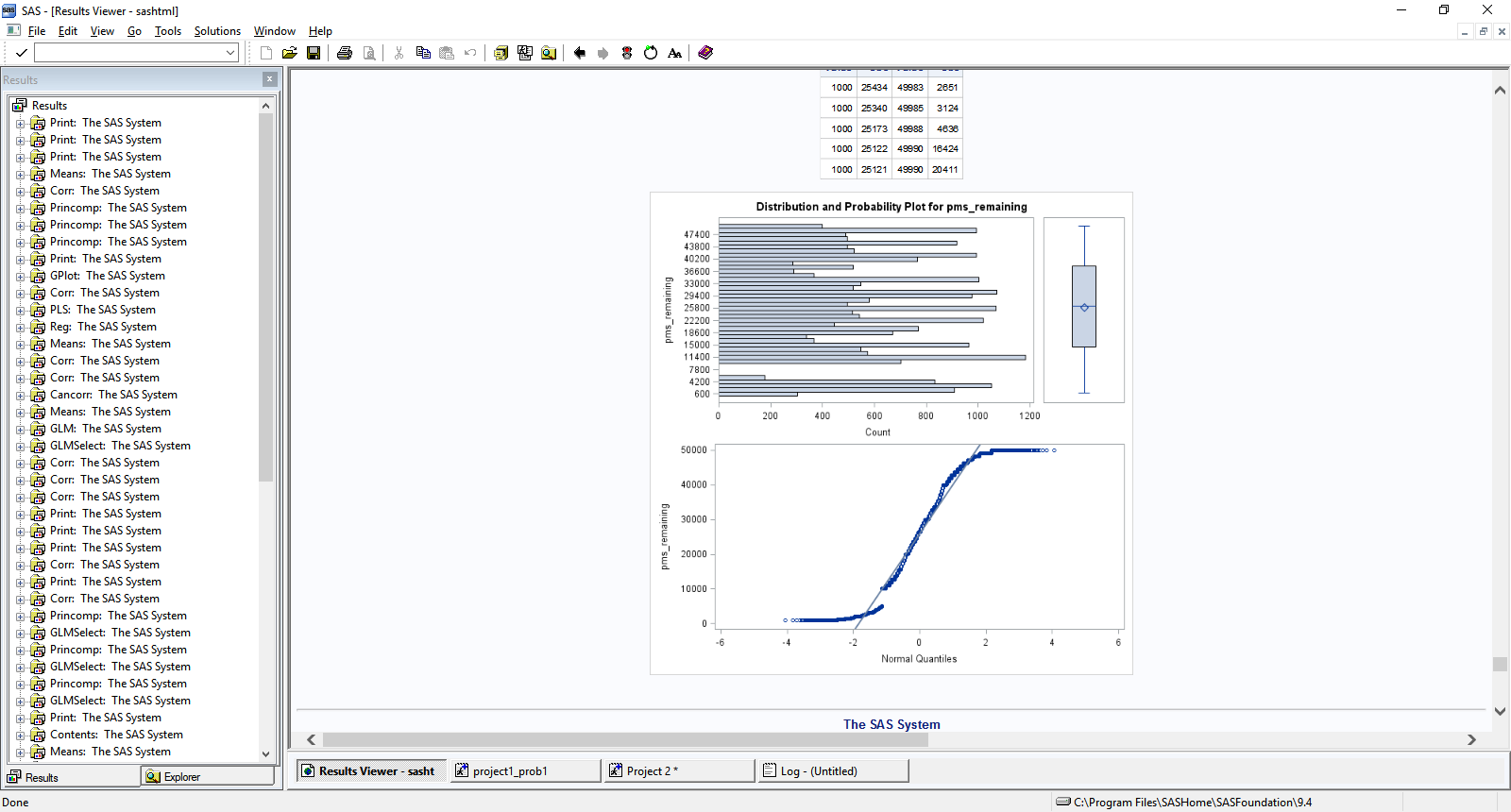
1. When “season” was imported, it contained 6182 missing values. The variable is formatted as a YYYY-YY value, denoting the Basketball season (for example, “2001-02”, for the 2001-2002 season. When imported, SAS automatically formats this fiels as a DateTime, but assumes that the “-YY” value is a month. So it interprets any “-YY” value greater than 12 as an error (for example, the “2012-13” season), and inputs a null value. We corrected this by removing the hyphen and reformatting the value as YYYYYY.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **N** | **N Miss** | **Notes** |
| recId | 25697 | 0 |  |
| game\_event\_id | 25697 | 0 |  |
| game\_id | 25697 | 0 |  |
| lat | 25697 | 0 |  |
| loc\_x | 25697 | 0 |  |
| loc\_y | 25697 | 0 |  |
| lon | 25697 | 0 |  |
| minutes\_remaining | 25697 | 0 |  |
| period | 25697 | 0 |  |
| playoffs | 25697 | 0 |  |
| season | 19515 | 6182 | (After reformatting, removed hyphen, no missing values) |
| seconds\_remaining | 25697 | 0 |  |
| shot\_distance | 25697 | 0 |  |
| shot\_made\_flag | 25697 | 0 |  |
| team\_id | 25697 | 0 |  |
| game\_date | 25697 | 0 |  |
| shot\_id | 25697 | 0 |  |
| attendance | 25697 | 0 |  |
| arena\_temp | 25697 | 0 |  |
| avgnoisedb | 25697 | 0 |  |

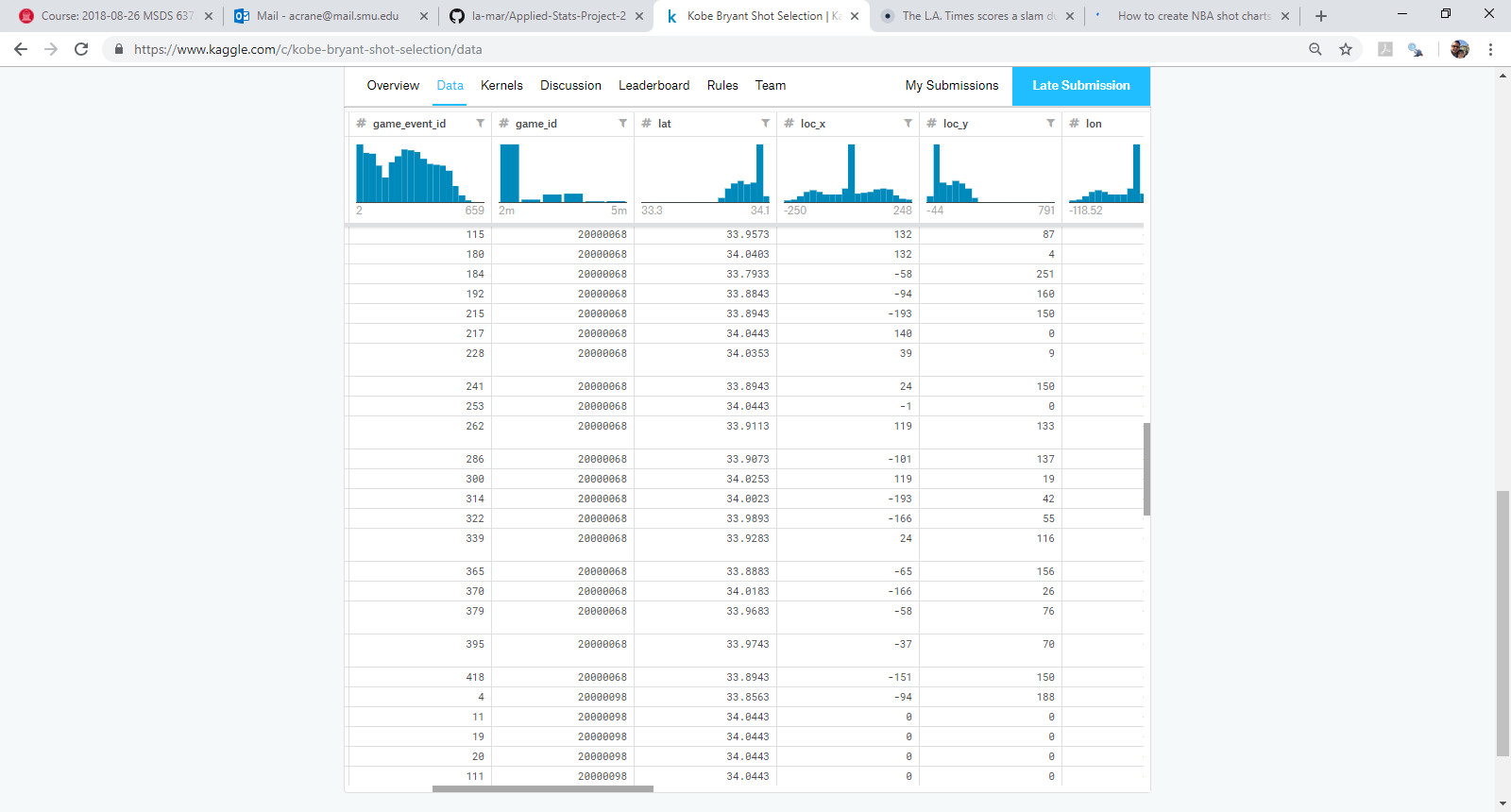
1. The fields “period”, “minutes\_remaining”, and “seconds\_remaining” together form a point in time of the shot attempt. The problem with this arrangement is repeating values of minutes (by period) and seconds (by periods and minutes). By concatenating these fields, we introduce a new variable that can have its own distribution, one we hope that is more meaningful than the former three.

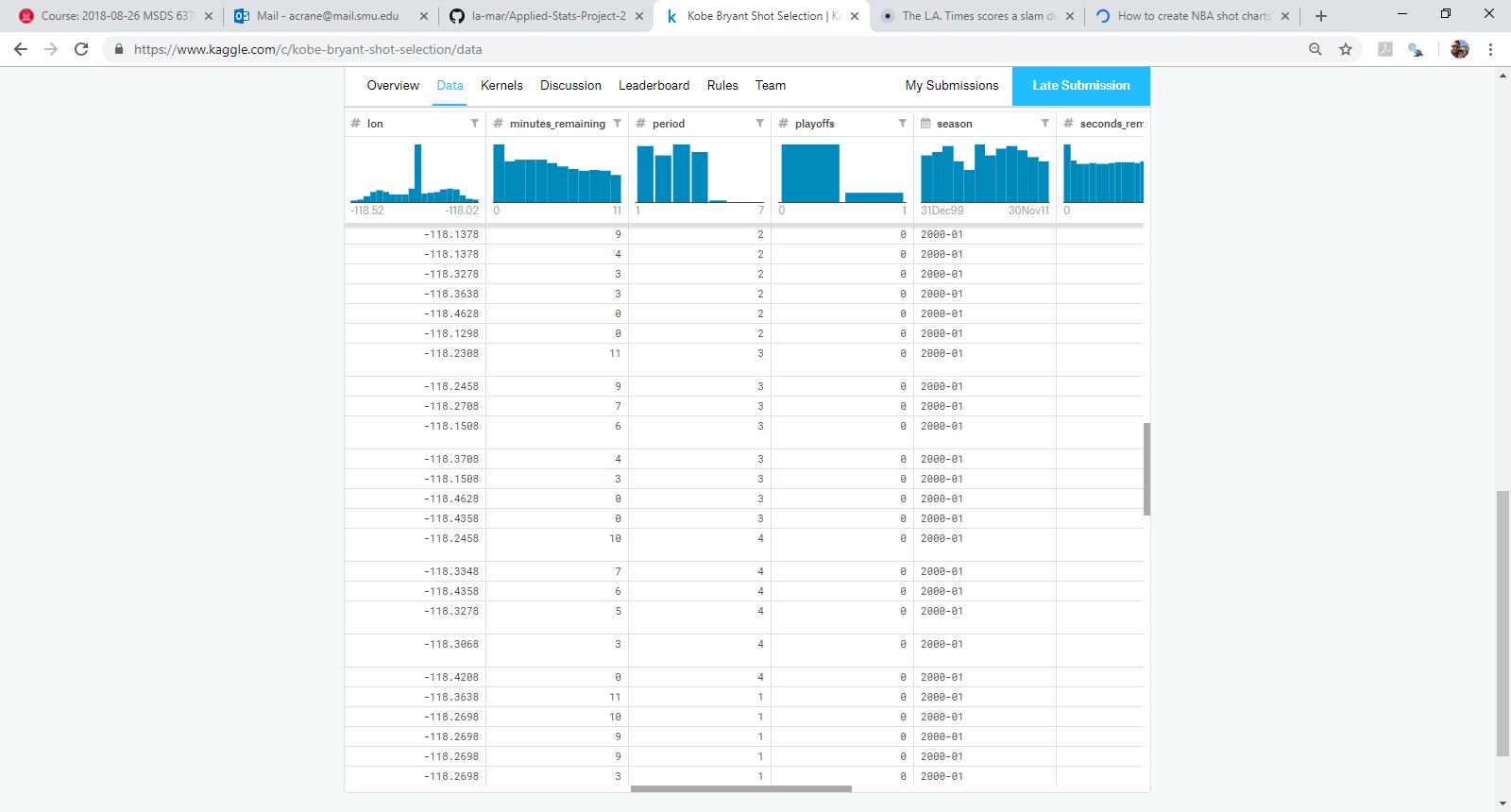


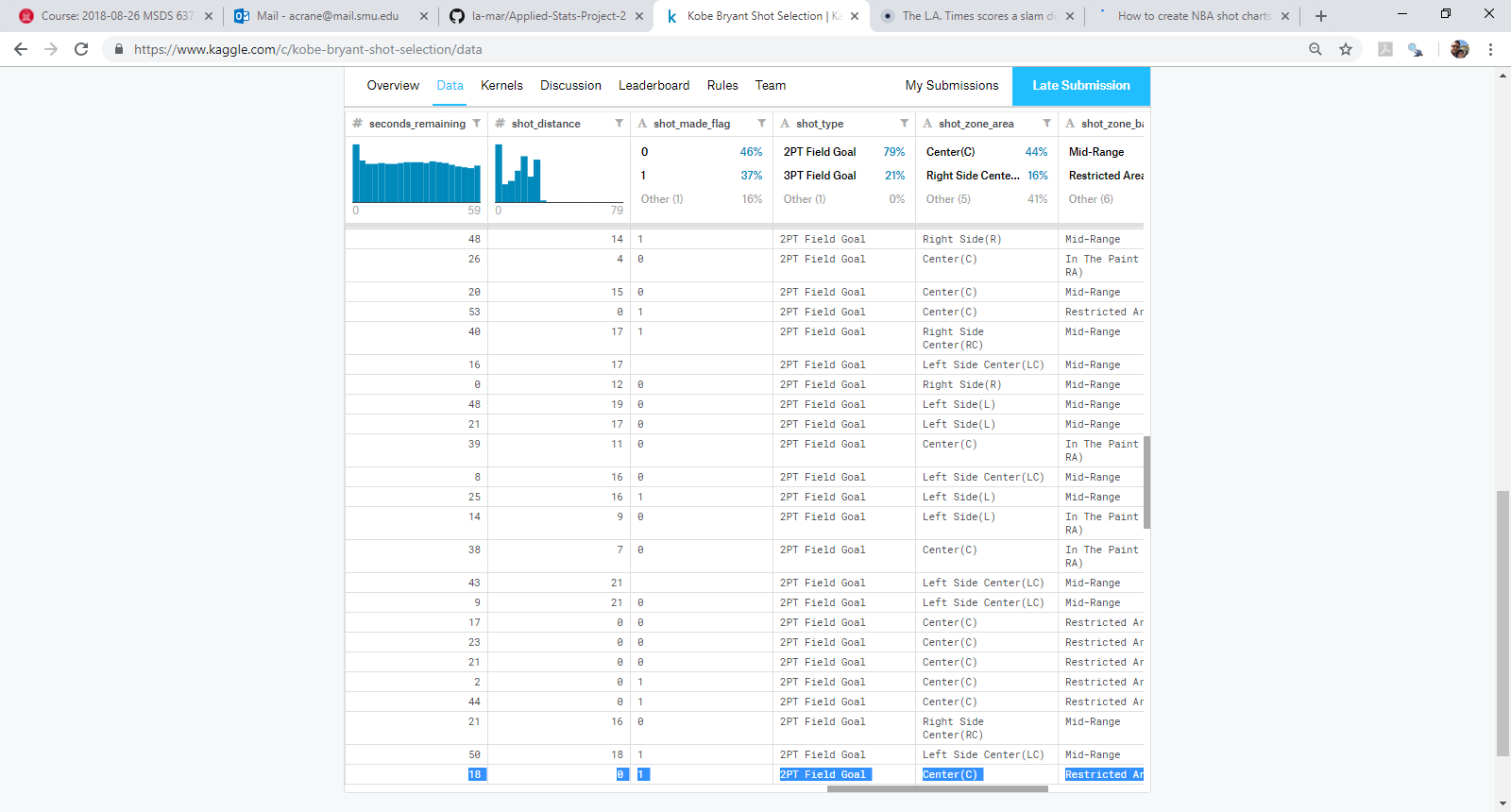




* **Address and identify outliers**







**Notes on outliers:**

Game\_event\_id Identification variable

Game\_id Identification variable

Lat and loc\_y Appear to be the same data (y-axis), based on location on the court. The tall bar indicates the frequency of shots taken at the location near the basket

Loc\_x and lon Appear to be the same data (x-axis), based on location on the court. The tall bar indicates the frequency of shots taken at the location near the basket

Minutes\_remaining Suggests that there is evidence that Kobe took more shots as the period progressed (higher shot counts closer to the end of the period)

Period Indicates similar shot frequency across periods

Playoffs Playoffs were a rare event in the season, hence the lower shot frequency overall

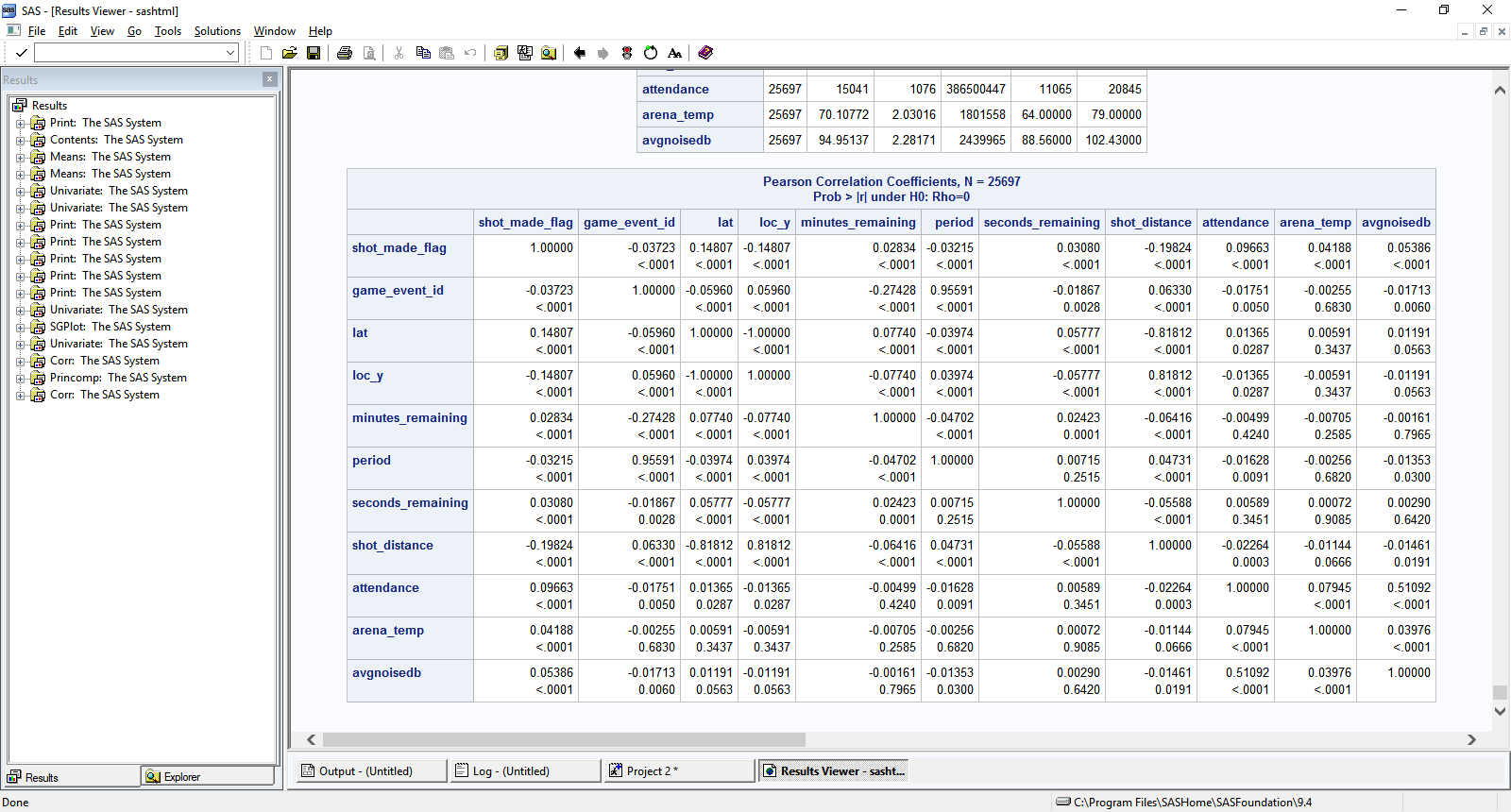
Season Shot frequency was lower in 2 seasons

Seconds\_remaining Suggests that there is evidence that Kobe took more shots in the final seconds (higher shot counts closer to the end of the last minute in the period, correlates with shot frequency in the minutes\_remaining variable above)

Shot\_distance High shot frequency at the basket, at the top of the key (field goals), and the three-point line. Shot frequency was comparatively rarer beyond the three-point line

We attempted to do some log transformation of some variables, but they did not appear to make a difference in our models, so we left them in, untransformed.

* **Address any Multicollinearity Issues**



**Notes on multicollinearity:**

After subsetting on the variables that had a high correlation with *shot\_made\_flag* (p < 0.0001), we took these variables and ran a correlation matrix on them. As you can see in the above table, *period* has evidence of multicollinearity with *game\_event\_id*, and *shot\_distance* has evidence of multicollinearity with *lat* and *loc\_y*. We will therefore drop *game\_event\_id*, *lat*, and *loc\_y* and continue using *period* and *shot\_distance* in our analysis.

Our highest performing model (Model 6) was based on the following reduced variable list:

Shot\_distance

Period

Minutes\_remaining

Seconds\_remaining

Attendance

Arena\_temp

Avgnoisedb

**Build models to provide arguments and evidence for or against the propositions below:**

We built a total of 7 models to answer these questions and compare for the most accurate answer. As you can see, the models evolved as we became more familiar with the data.

Model 1. GLM using PCA

Model 2. Logistic using PCA

Model 3. Logistic using non-PCA (shot distance to predict shot made)

Model 4A. Logistic using non-PCA (shot distance to predict shot made, using playoffs data set)

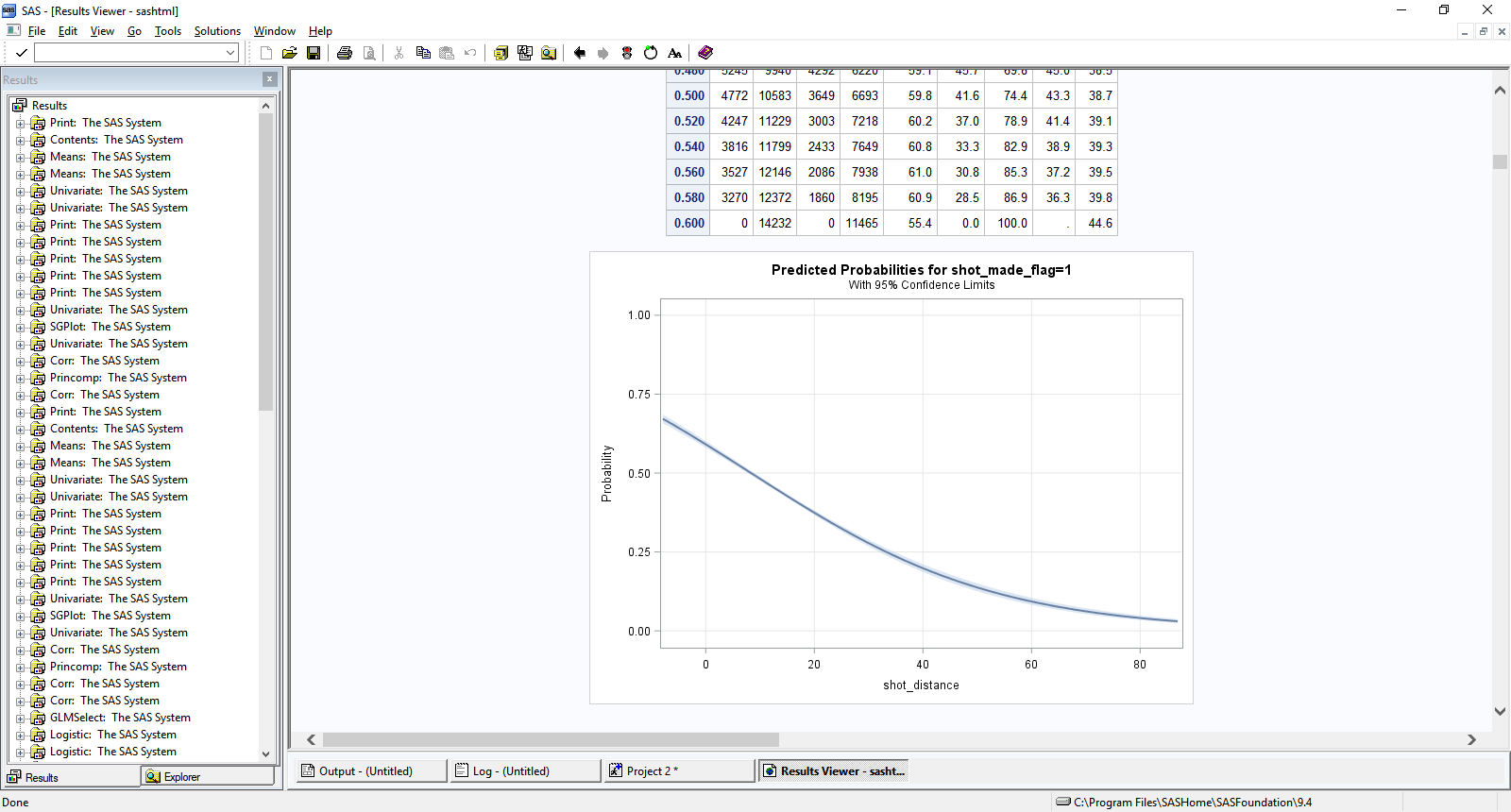
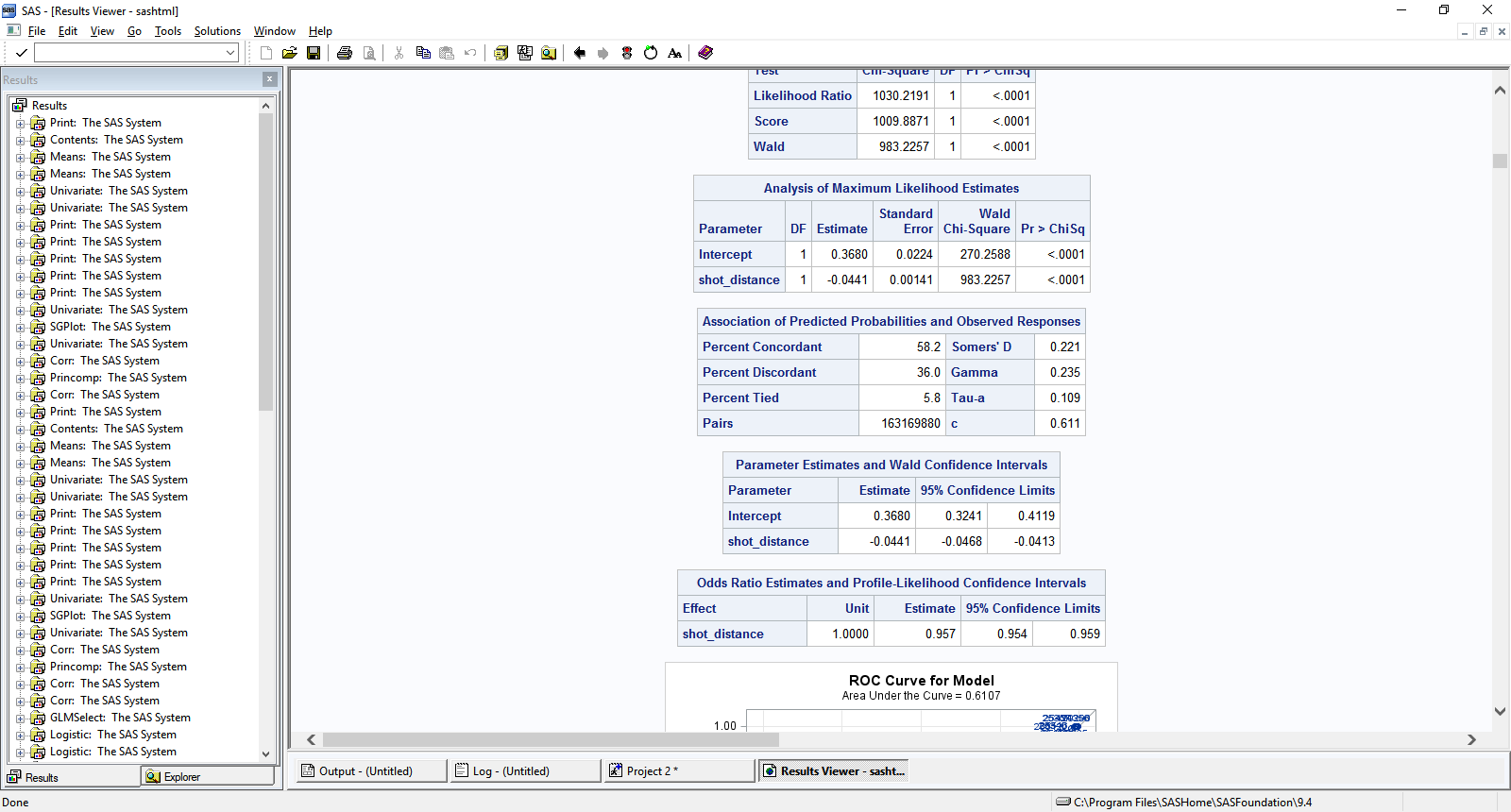
Model 4B. Logistic using non-PCA (shot distance to predict shot made, using not at playoffs data set)

Model 5. Logistic using non-PCA (mix of variables selected from PCA matrix to predict shot made)

Model 6. Logistic using non-PCA (final mix of variables selected from CORR matrix [after removing collinear variables] to predict shot made)

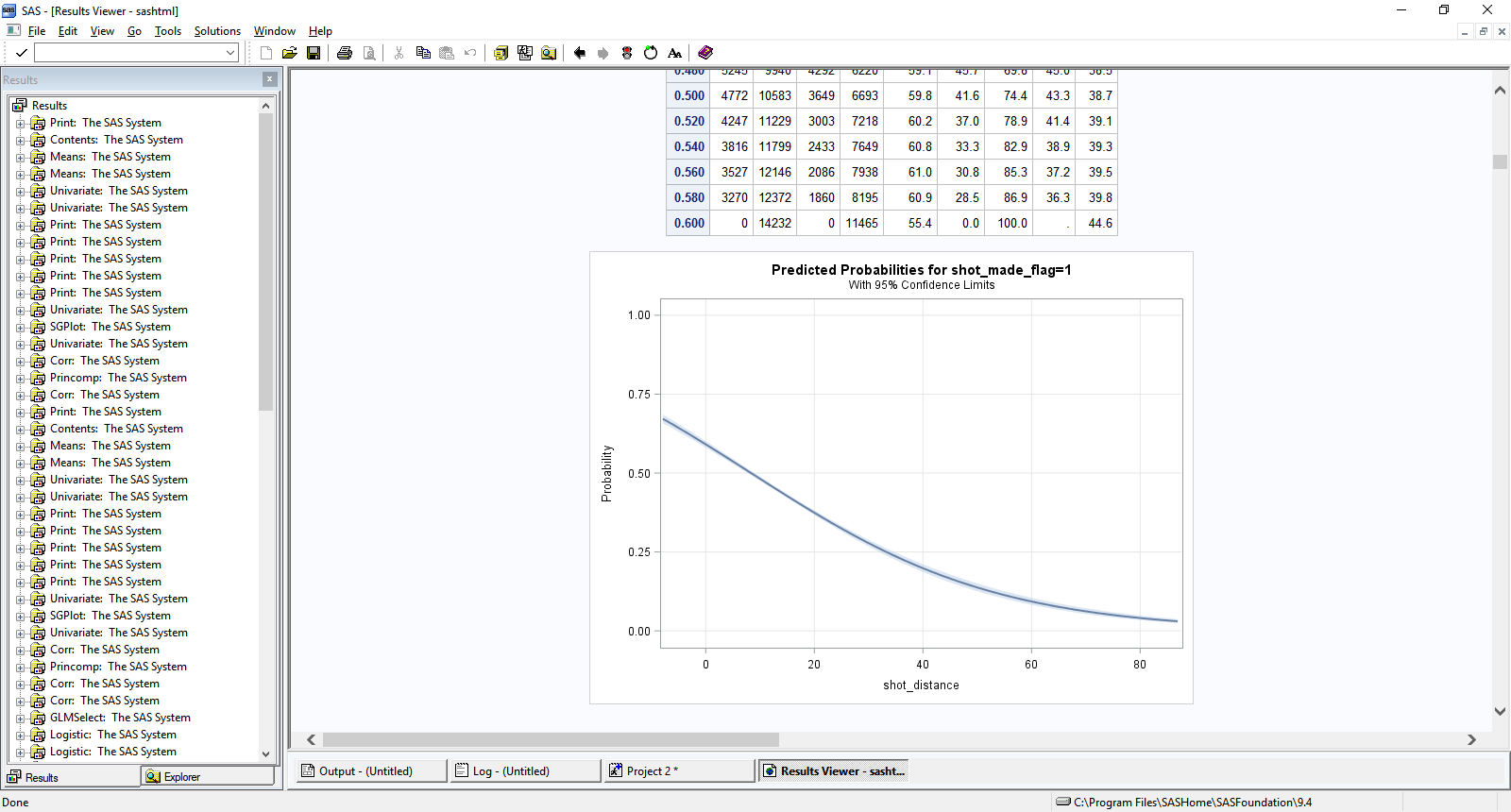
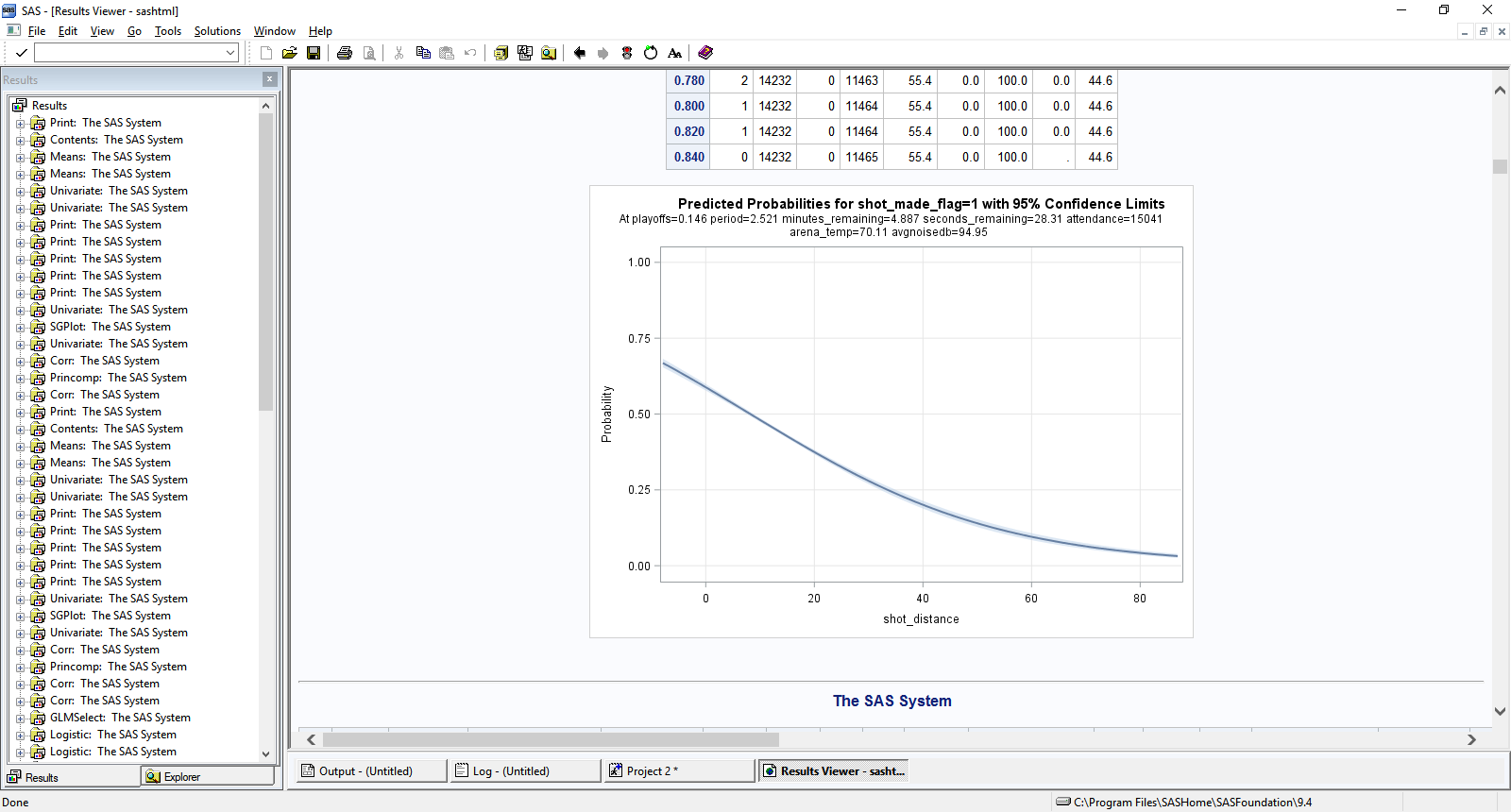
Model 7. Linear Discriminant Analysis (final mix of variables selected from CORR matrix [after removing collinear variables] to predict shot made)

* **The odds of Kobe making a shot decrease with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)**

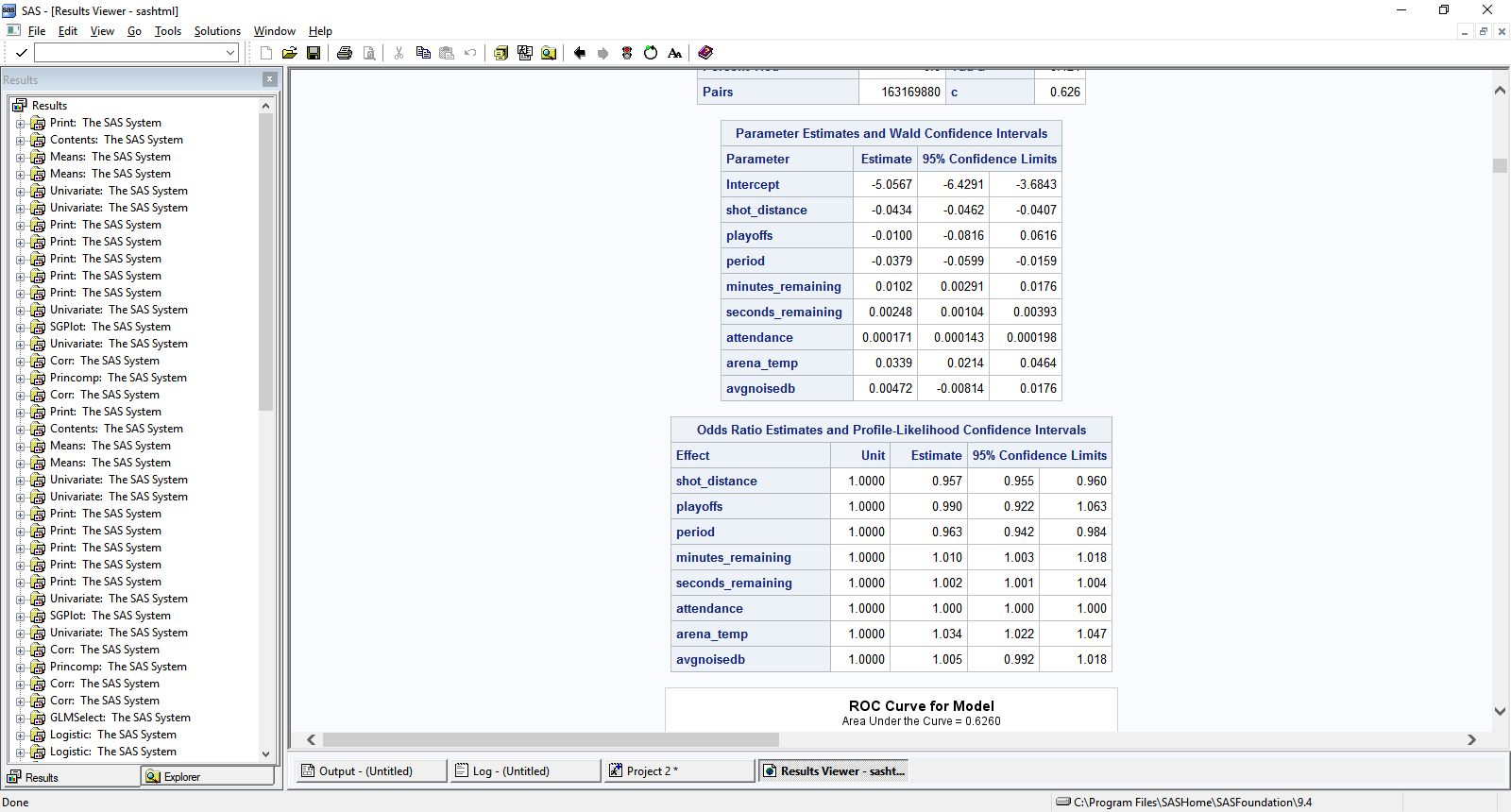


Using Model 3 (using only shot\_distance to predict shot\_made\_flag), we see strong evidence (p < 0.0001) that shot\_distance is significant, in a negative direction. In other words, for every foot further away from the basket, Kobe was 4.4% less likely to make the shot.

* **The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop. If there is evidence of this, quantify this relationship. (CIs, plots, etc.)**

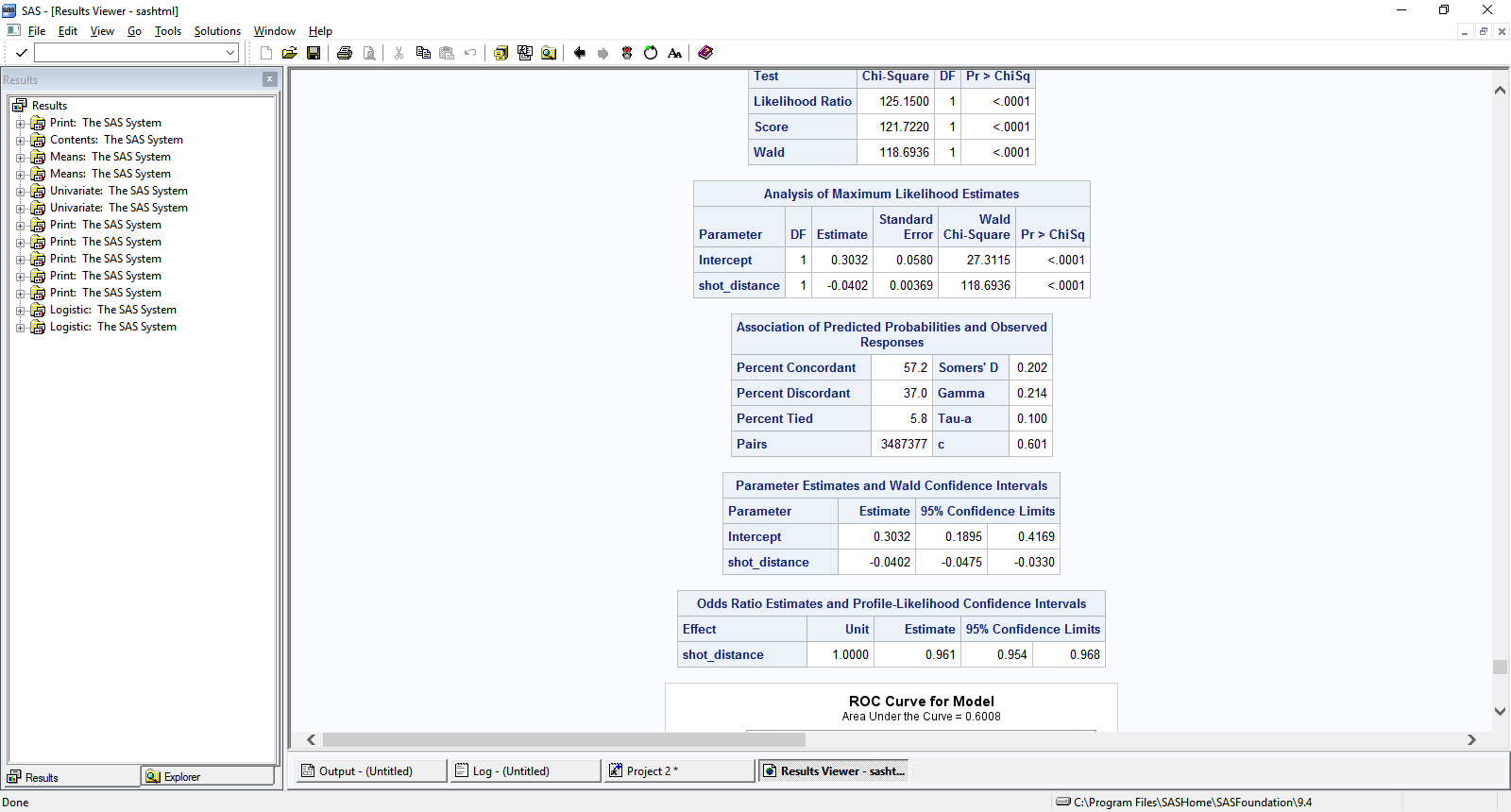


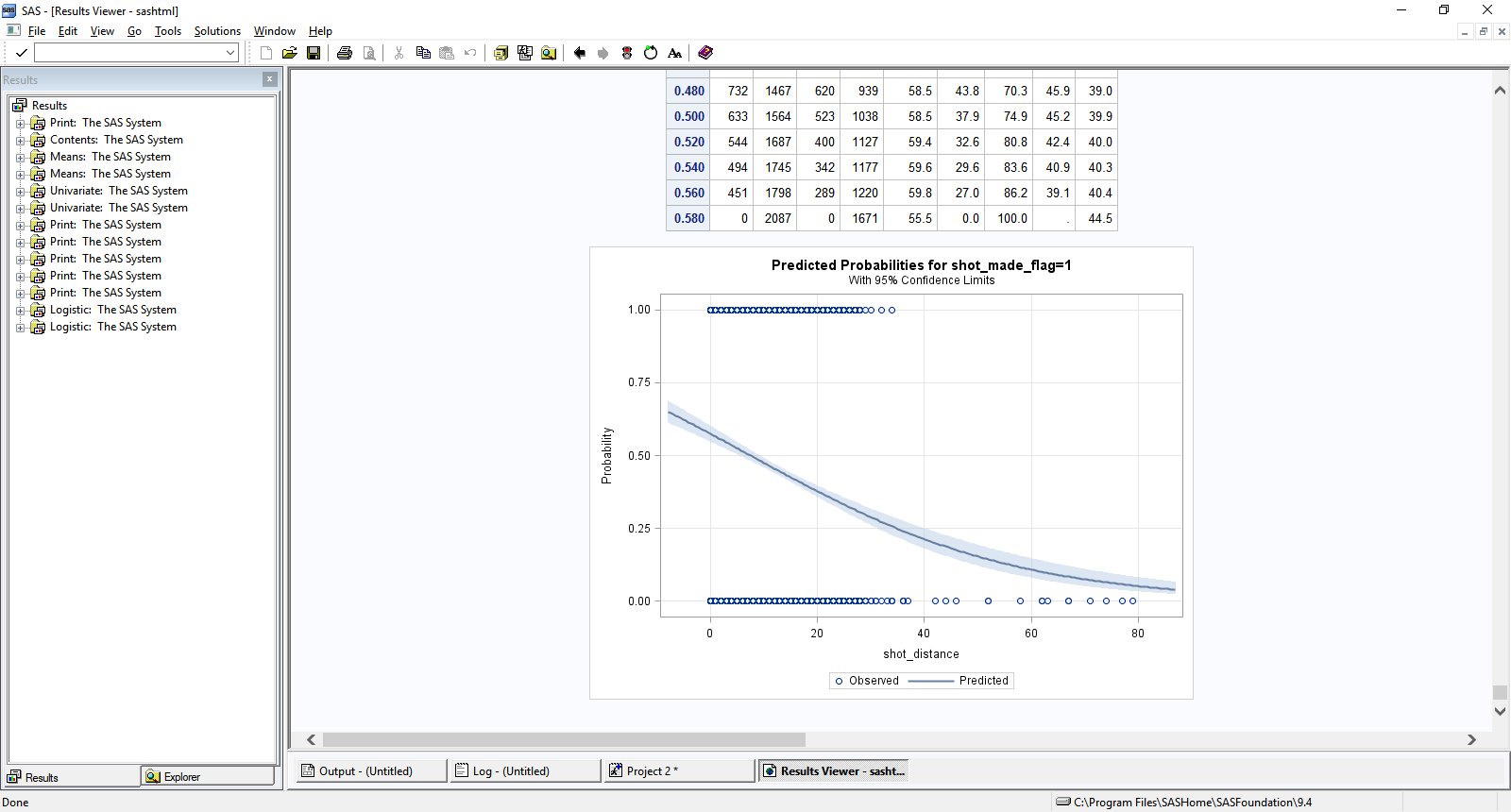
In every model we used (example models 3 and 6 above), we see a near linear relationship of distance to shot made up to about 20-25 feet away. Even with all these additional variables, we see that for every foot further away from the basket, Kobe was 4.3% less likely to make the shot, consistent with the above findings. Beyond those distances, the relationship becomes non-linear, as it curves out to longer distances up to 80 feet and beyond. Parameter estimates from Model 6 are shown below:



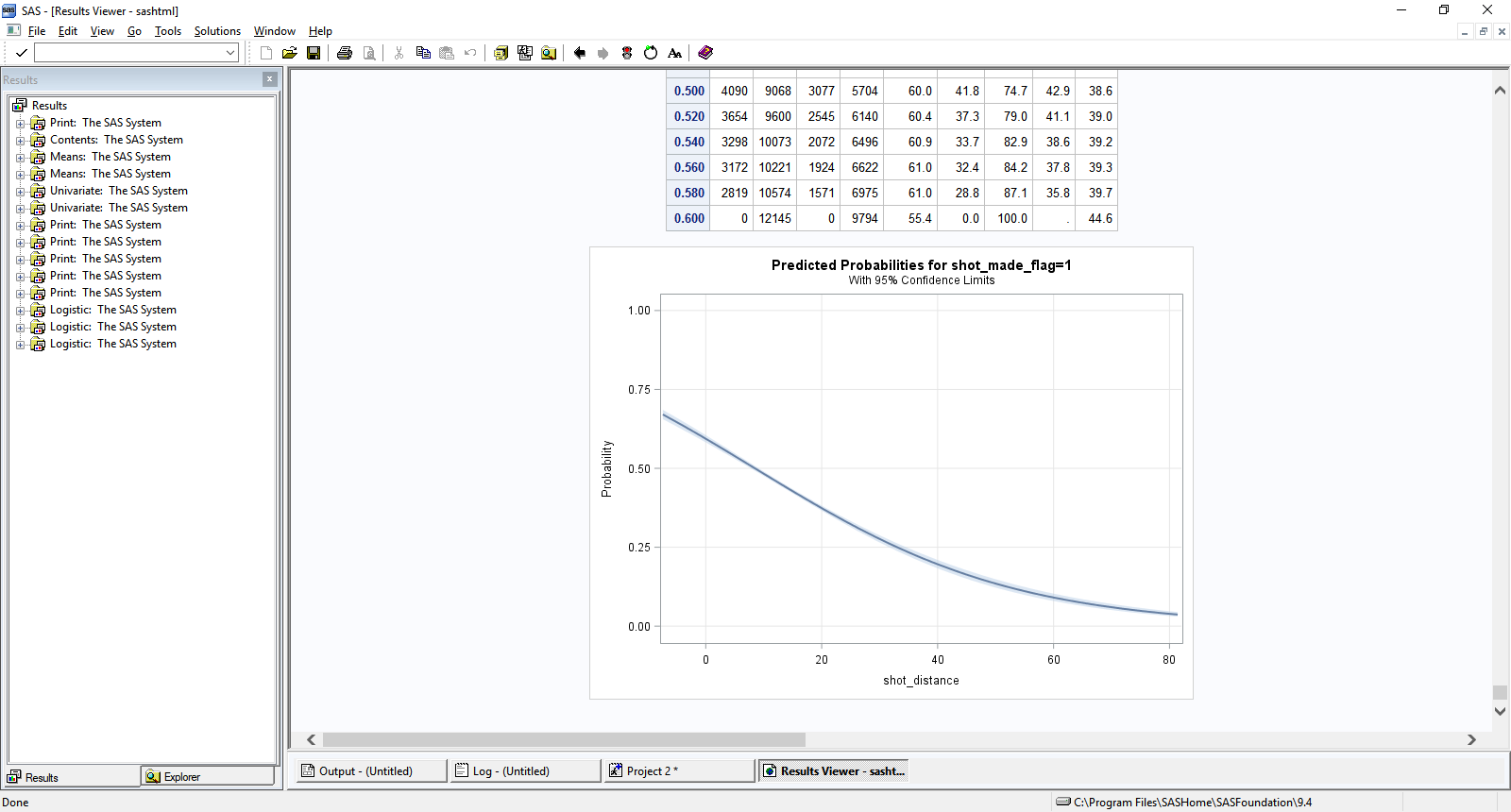
* **The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs. Quantify your findings with statistical evidence one way or the other. (Tests, CIs, plots, etc.)**

We first split the data into two data sets, playoffs and no playoffs. Then, using Model 4 (shot\_distance), we ran this against each data set (playoffs and non playoffs).

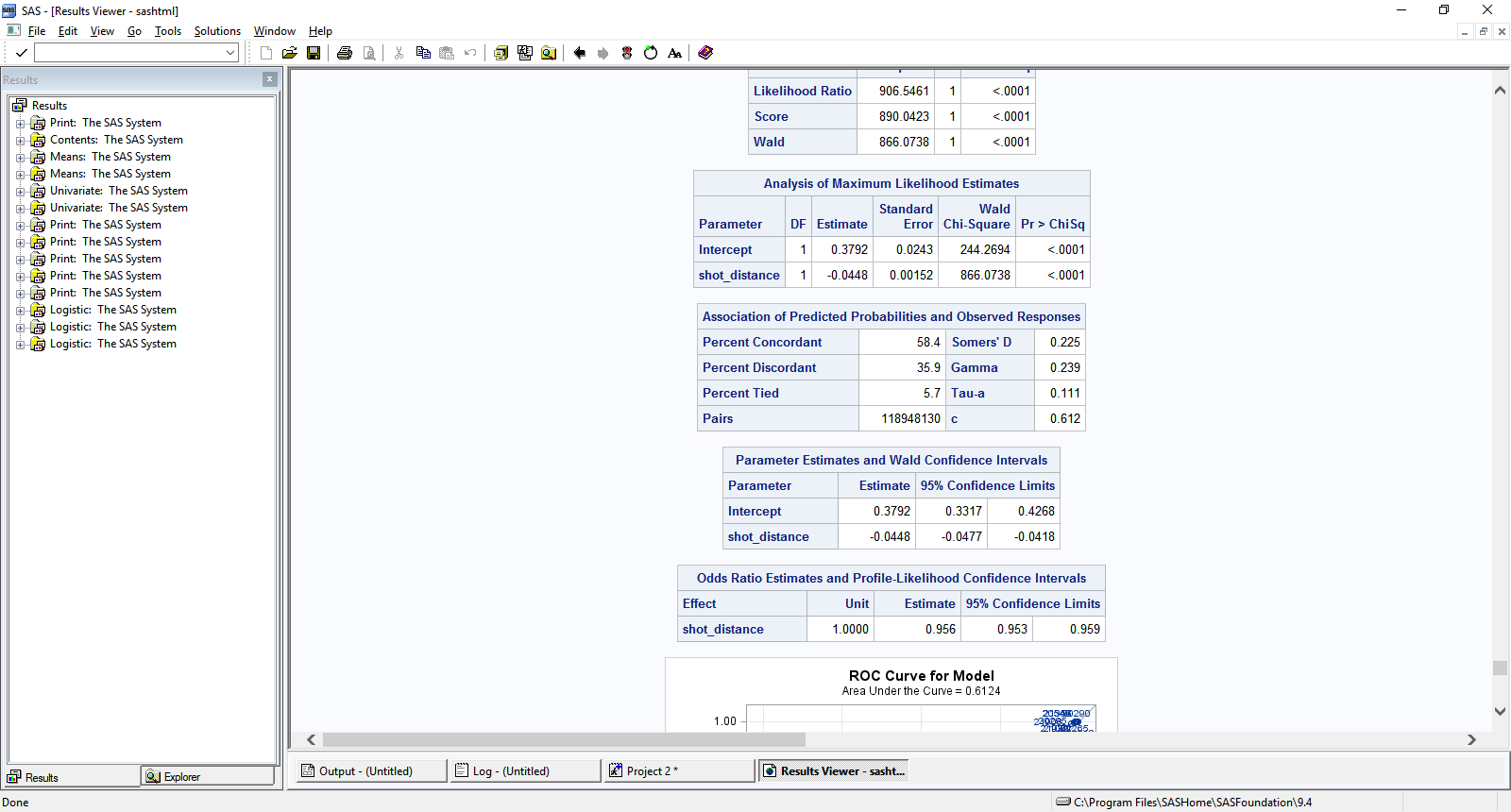




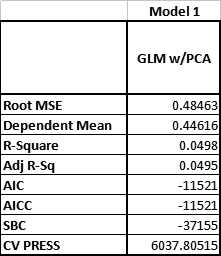
**In the playoffs, we see that for every foot further away from the basket, Kobe was only 4% less likely to make the shot, an improvement of nearly half of a full percentage point over the average.**

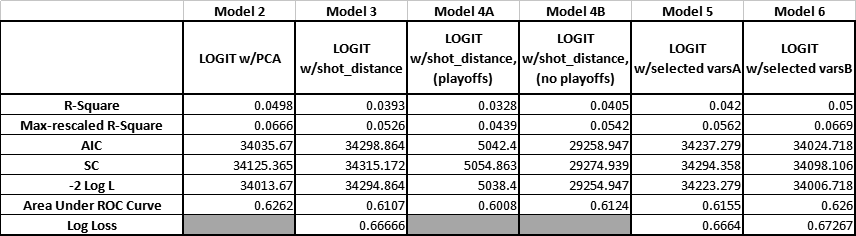


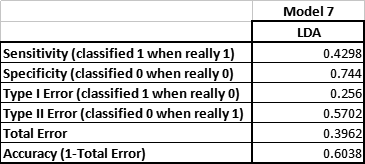
**When not in the playoffs, we see that for every foot further away from the basket, Kobe was 4.5% less likely to make the shot, consistent with the above results, but nearly half of a full percentage point worse than when he performed in the playoffs.**



**Appendix A. Model comparisons**







Note: Log loss scores calculated via Kaggle submission, which used log loss as scoring criteria (https://www.kaggle.com/c/kobe-bryant-shot-selection#evaluation). According to our model log loss scores, Model 5 performed the best, with a result of 0.6664. Model variables of Model 5 were:

Shot\_distance

Playoffs

Arena\_temp

Game\_event\_id

lat

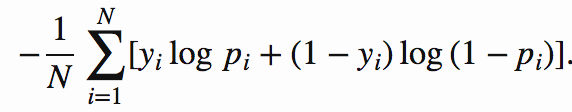
lon

**FORMAT OF THE PAPER - CONTINUED**

Build a predictive model to classify shots as missed or made. You should produce at least 1 of each type of model:

* A logistic regression model.
* A Linear Discriminant Analysis (LDA) model.

Evaluation: Compare each competing models with the AUC, Mis-Classification Rate, Sensitivity, Specificity and objective / loss function. The log loss function of the model should be used to assess the model fit:



Where N is the total number classifications, yi is the shot\_made\_flag and pi is the probability from the model of each outcome (shot made or shot missed.)

**ASSESSMENT / EVALUATION:**

Good papers traditionally have the following characteristics:

1. They are presented in an organized, neat and consistent fashion. (Labeled plots, figures and tables, consistently formatted, indented and labeled headers and sub headers, etc.) Given that each group has 3 members, the paper should only have one look and feel. Titles, headers, sub headers, figures, tables, etc. should all look the same and have numbering that is consistent.
2. There are no typos, misspelled words, grammatical mistakes, etc.
3. They use a variety of methods.
4. Creative methods are used.
5. They have input from all group members and are developed iteratively over time as opposed to all at once such as the night before.

The group with the lowest log loss score will be awarded an additional 3 points for the project.

**SOFTWARE AND METHODS:**

You may use any software and must use only the methods we have studied thus far in the course. That being said, you can use innovative techniques inside of those methods like model averaging, cross validation or creating new variables from the ones in the data set. If you have any questions about this please let me know and we can discuss your ideas.

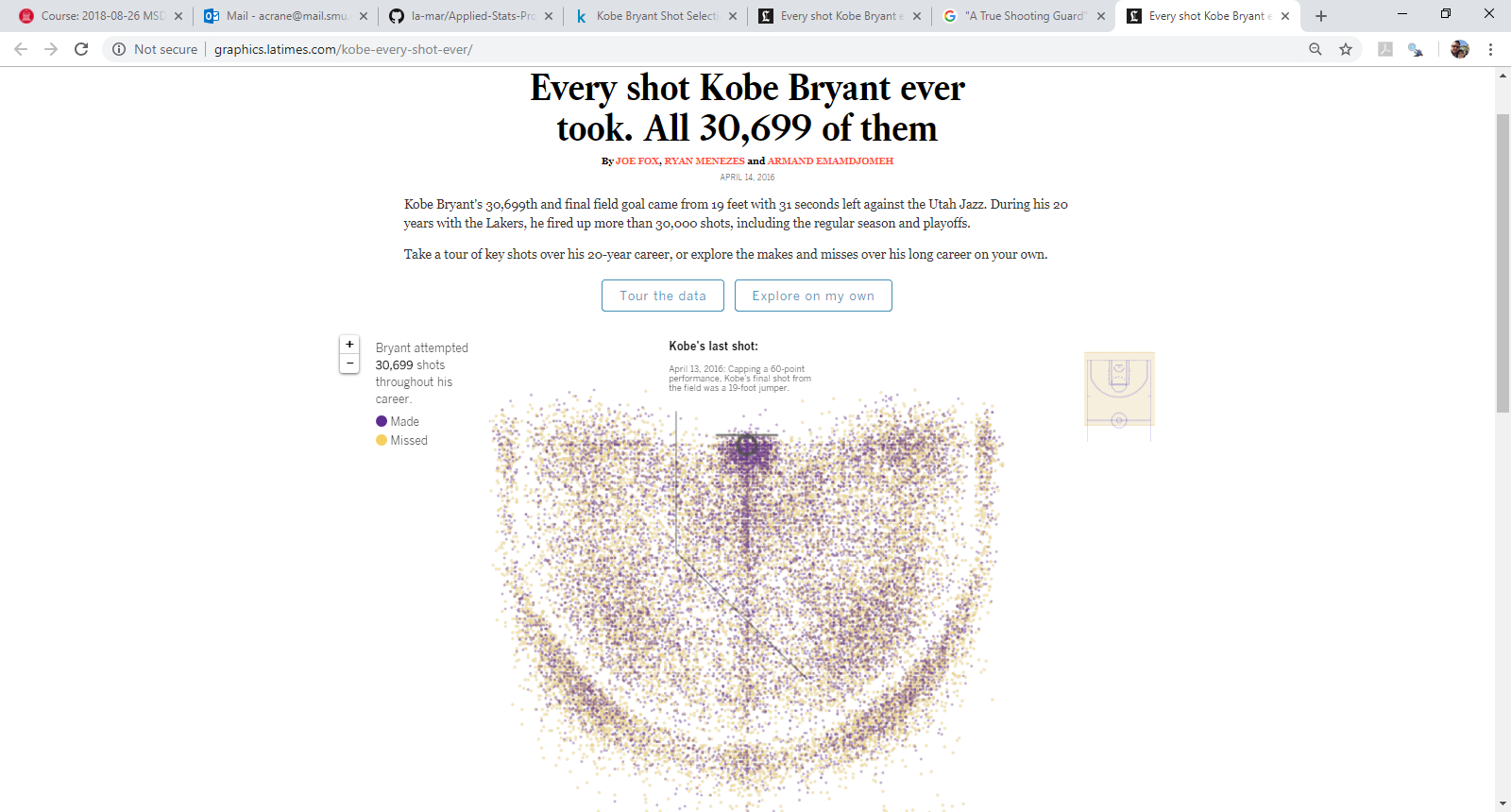
**About Kobe Bryant**

**Kobe Bean Bryant** (born August 23, 1978) is an American former professional [basketball](https://en.wikipedia.org/wiki/Basketball" \o "Basketball) player. He played his entire 20-year career with the [Los Angeles Lakers](https://en.wikipedia.org/wiki/Los_Angeles_Lakers" \o "Los Angeles Lakers) of the [National Basketball Association](https://en.wikipedia.org/wiki/National_Basketball_Association" \o "National Basketball Association) (NBA). He entered the NBA directly from high school and won five [NBA championships](https://en.wikipedia.org/wiki/NBA_Finals" \o "NBA Finals) with the Lakers. Bryant is an 18-time [All-Star](https://en.wikipedia.org/wiki/NBA_All-Star" \o "NBA All-Star), 15-time member of the [All-NBA Team](https://en.wikipedia.org/wiki/All-NBA_Team" \o "All-NBA Team), and 12-time member of the [All-Defensive team](https://en.wikipedia.org/wiki/NBA_All-Defensive_Team" \o "NBA All-Defensive Team). He [led the NBA in scoring](https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_annual_scoring_leaders" \o "List of National Basketball Association annual scoring leaders) during two seasons and ranks third on the league's [all-time regular season scoring](https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_career_scoring_leaders" \o "List of National Basketball Association career scoring leaders) and fourth on the [all-time postseason scoring](https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_career_playoff_scoring_leaders" \o "List of National Basketball Association career playoff scoring leaders) list. He holds the NBA record for the [most seasons playing with one franchise for an entire career](https://en.wikipedia.org/wiki/List_of_NBA_players_who_have_spent_their_entire_career_with_one_franchise" \o "List of NBA players who have spent their entire career with one franchise) and is widely regarded as one of the greatest basketball players of all time.[[3]](https://en.wikipedia.org/wiki/Kobe_Bryant" \l "cite_note-4)[[4]](https://en.wikipedia.org/wiki/Kobe_Bryant" \l "cite_note-5)[[5]](https://en.wikipedia.org/wiki/Kobe_Bryant" \l "cite_note-6)[[6]](https://en.wikipedia.org/wiki/Kobe_Bryant" \l "cite_note-7) Bryant is the first [guard](https://en.wikipedia.org/wiki/Guard_(basketball)" \o "Guard (basketball)) in NBA history to play at least [20 seasons](https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_seasons_played_leaders" \o "List of National Basketball Association seasons played leaders).

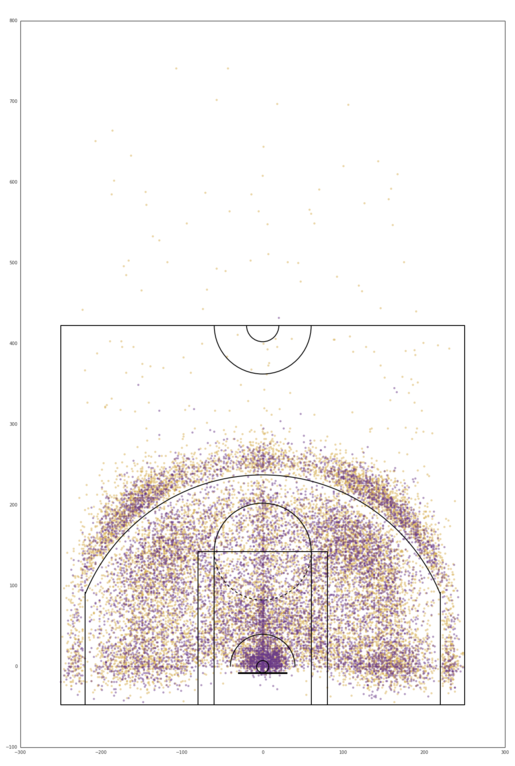
The son of former NBA player [Joe Bryant](https://en.wikipedia.org/wiki/Joe_Bryant" \o "Joe Bryant), Kobe Bryant enjoyed a successful high school basketball career at [Lower Merion High School](https://en.wikipedia.org/wiki/Lower_Merion_High_School" \o "Lower Merion High School) in [Pennsylvania](https://en.wikipedia.org/wiki/Pennsylvania" \o "Pennsylvania), where he was recognized as the top high school basketball player in the country. He declared for the [NBA draft](https://en.wikipedia.org/wiki/NBA_draft" \o "NBA draft) upon graduation and was selected in the 13th overall pick in the [1996 NBA draft](https://en.wikipedia.org/wiki/1996_NBA_draft" \o "1996 NBA draft) by the [Charlotte Hornets](https://en.wikipedia.org/wiki/1996%E2%80%9397_Charlotte_Hornets_season" \o "1996–97 Charlotte Hornets season), who traded him to the Lakers. As a rookie, Bryant earned himself a reputation as a high-flyer and a fan favorite by winning the 1997 [Slam Dunk Contest](https://en.wikipedia.org/wiki/Slam_Dunk_Contest" \o "Slam Dunk Contest), and he was named an All-Star by his second season. Despite a [feud between them](https://en.wikipedia.org/wiki/Shaq%E2%80%93Kobe_feud" \o "Shaq–Kobe feud), Bryant and [Shaquille O'Neal](https://en.wikipedia.org/wiki/Shaquille_O'Neal" \o "Shaquille O'Neal) led the Lakers to three consecutive NBA championships from [2000](https://en.wikipedia.org/wiki/2000_NBA_Finals" \o "2000 NBA Finals) to [2002](https://en.wikipedia.org/wiki/2002_NBA_Finals" \o "2002 NBA Finals).

In 2003, Bryant was [accused of sexual assault](https://en.wikipedia.org/wiki/Kobe_Bryant_sexual_assault_case" \o "Kobe Bryant sexual assault case) in [Colorado](https://en.wikipedia.org/wiki/Colorado" \o "Colorado), but the charges were eventually dropped, and a civil suit was settled out of court. After the Lakers lost the [2004 NBA Finals](https://en.wikipedia.org/wiki/2004_NBA_Finals" \o "2004 NBA Finals), O'Neal was traded to the [Miami Heat](https://en.wikipedia.org/wiki/Miami_Heat" \o "Miami Heat). Bryant became the cornerstone of the Lakers, and he led the NBA in scoring during the [2005–06](https://en.wikipedia.org/wiki/2005%E2%80%9306_NBA_season" \o "2005–06 NBA season) and [2006–07](https://en.wikipedia.org/wiki/2006%E2%80%9307_NBA_season" \o "2006–07 NBA season) seasons. In 2006, he scored a career-high 81 points against the [Toronto Raptors](https://en.wikipedia.org/wiki/2005%E2%80%9306_Toronto_Raptors_season" \o "2005–06 Toronto Raptors season), the second [most points scored in a single game](https://en.wikipedia.org/wiki/List_of_National_Basketball_Association_single-game_scoring_leaders" \o "List of National Basketball Association single-game scoring leaders) in league history behind [Wilt Chamberlain's 100-point game](https://en.wikipedia.org/wiki/Wilt_Chamberlain's_100-point_game" \o "Wilt Chamberlain's 100-point game) in 1962. Bryant was awarded the regular season's [Most Valuable Player Award](https://en.wikipedia.org/wiki/NBA_Most_Valuable_Player_Award" \o "NBA Most Valuable Player Award) (MVP) in 2008. After losing in the [2008 NBA Finals](https://en.wikipedia.org/wiki/2008_NBA_Finals" \o "2008 NBA Finals), he led the Lakers to two consecutive championships in [2009](https://en.wikipedia.org/wiki/2009_NBA_Finals" \o "2009 NBA Finals) and [2010](https://en.wikipedia.org/wiki/2010_NBA_Finals" \o "2010 NBA Finals), earning the [Finals MVP Award](https://en.wikipedia.org/wiki/Bill_Russell_NBA_Finals_Most_Valuable_Player_Award" \o "Bill Russell NBA Finals Most Valuable Player Award) on both occasions. He continued to be among the top players in the league through 2013 when the 34-year-old Bryant suffered a [torn Achilles tendon](https://en.wikipedia.org/wiki/Torn_Achilles_tendon" \o "Torn Achilles tendon). Although he recovered, his play was limited the following two years by season-ending injuries to his knee and shoulder, respectively. Citing his physical decline, he announced that he would retire after the [2015–16](https://en.wikipedia.org/wiki/2015%E2%80%9316_NBA_season" \o "2015–16 NBA season) season.

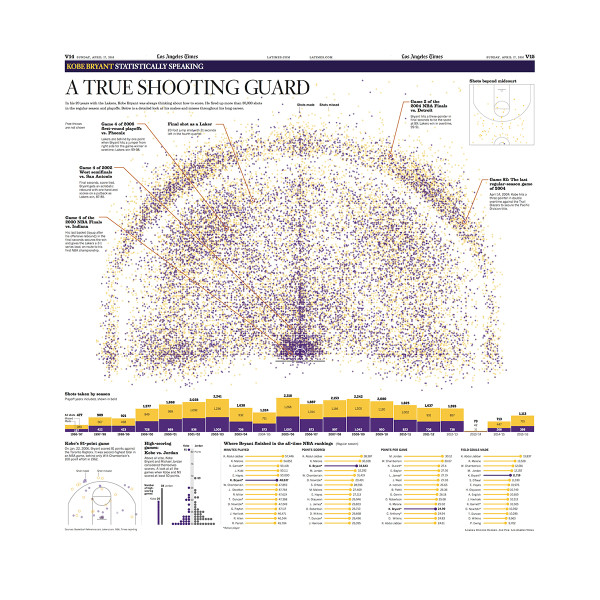
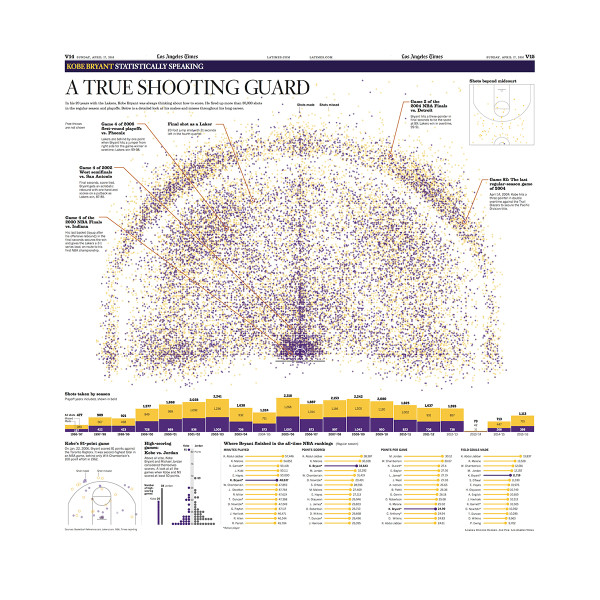
At 34 years and 104 days of age, Bryant became the youngest player in league history to reach 30,000 career points. He became the all-time leading scorer in Lakers franchise history on February 1, 2010, when he surpassed [Jerry West](https://en.wikipedia.org/wiki/Jerry_West" \o "Jerry West). During his third year in the league, Bryant was chosen to start the All-Star Game, and he would continue to be selected to start that game for a record 18 consecutive appearances until his retirement. His four [All-Star MVP Awards](https://en.wikipedia.org/wiki/NBA_All-Star_Game_Most_Valuable_Player_Award" \o "NBA All-Star Game Most Valuable Player Award) are tied for the most in NBA history. At the [2008](https://en.wikipedia.org/wiki/Basketball_at_the_2008_Summer_Olympics" \o "Basketball at the 2008 Summer Olympics) and [2012 Summer Olympics](https://en.wikipedia.org/wiki/Basketball_at_the_2012_Summer_Olympics" \o "Basketball at the 2012 Summer Olympics), he won [gold medals](https://en.wikipedia.org/wiki/Gold_medal" \o "Gold medal) as a member of the [U.S. national team](https://en.wikipedia.org/wiki/United_States_men's_national_basketball_team" \o "United States men's national basketball team). *[Sporting News](https://en.wikipedia.org/wiki/Sporting_News" \o "Sporting News)* and [TNT](https://en.wikipedia.org/wiki/TNT_(U.S._TV_network)" \o "TNT (U.S. TV network)) named Bryant the top NBA player of the 2000s. In 2018, Bryant won the [Academy Award for Best Animated Short Film](https://en.wikipedia.org/wiki/Academy_Award_for_Best_Animated_Short_Film" \o "Academy Award for Best Animated Short Film) for his film *[Dear Basketball](https://en.wikipedia.org/wiki/Dear_Basketball" \o "Dear Basketball)*.



Source: <http://graphics.latimes.com/kobe-every-shot-ever/>



***A note on the X Y grid locations on the court: The NBA data includes X and Y coordinates that refer to positions on the court. The units are tenths of a foot, so the sidelines are -250 and 250, because the court is 50 feet wide.***



*Source: Los Angeles Times, 4/17/2016, <https://store.latimes.com/product/7QAMLA157/a-true-shooting-guard-kobe-bryant-commemorative-sports-page>*